Performance of the Gold Standard and Machine Learning in Predicting Vehicle Transactions

Alina Lazar
Dept. of Computer Science and Information Systems
Youngstown State University
Youngstown, OH
alazar@ysu.edu

Ling Jin, Caitlin Brown, C. Anna Spurlock
Energy Technology Area
Lawrence Berkeley National Laboratory
Berkeley, CA
ljin,caitlinbrown,caspurlock@lbl.gov

Alexander Sim, Kesheng Wu
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA
asim, kwu@lbl.gov

***Abstract*—Logistic regression has long been the gold standard for choice modeling in the transportation field. Despite the rising popularity of machine learning (ML), few is applied to predicting the household vehicle transactions. To address the research gap, this paper presents a first use case of ML application to predicting household vehicle transaction decisions by leveraging a newly processed national panel data set. Model performances are reported for four ML models and the traditional multinomial logit model (MNL). Instead of treating the gold standard and ML models as competitors, this paper tries to use ML tools to inform the MNL model building process. We find the two gradient boosting based methods, CatBoost and LightGBM, are the best performing ML models; and improving logistic models with SHAP interpretation tools can achieve similar performance levels to the best performing ML methods.**

***Index Terms*—household vehicle transaction, multinomial logit, gradient boosting, SHAP values, treeExplainer**

I. Introduction

Methodologies for predicting household vehicle transaction choices are instrumental to longer term transportation planning and creating sustainable transportation systems. Logistic regression has long been the gold standard in choice modeling for transportation problems [1]. These choice models are based on random utility maximisation theory and the estimated coefficients can straightforwardly quantify feature effects as changes in odds ratios.

Unlike statistical models, like multinomial logit, that impose a predetermined structure, machine learning (ML) models rely on data-driven heuristics to arrive at their solutions. Although ML applications to predicting household vehicle transactions are rare, there are published studies on predicting other travel behaviors, such as transportation mode choices, that have shown that ML models outperform traditional statistical models [2], [3], [4]. However the limited interpretability of ML models have limited their wider adoption.

Rather than treating the gold standard and ML models as competitors, opportunities exist to marry the two. Recent advances in "Explainable AI" [5], [6] have improved the interpretability of tree-based models exploring high-dimensional feature space. Behavior insights from the ML models, such as individual feature importance and their interactions, have been proposed to be incorporated into the logistic model building process to improve its model specification and prediction performance [7], [8].

Despite the rising body of ML literature on travel behavior research (see review in [1]), few studies have addressed household vehicle transactions. Predicting the dynamics of vehicle transactions requires longitudinal data that are difficult to collect from the life courses of individual households. Current data collection are mostly reliant on cross-sectional surveys with small sample sizes that limit the application of data-driven ML models.

To address this research gap, this paper presents a first use-case of applying ML to prediction of household vehicle transaction decisions by leveraging a newly processed nationally representative panel data set. Our contribution to the travel behavior research includes: (1) the first study documenting the performance of various ML methods on predicting vehicle-level transactions; (2) comparison of the performance between ML and the logistic models with a comprehensive set of metrics; and (3) demonstrating performance improvements in the logistic model via SHAP interpretation tools incorporated into the model building process. We find the two gradient boosting based methods, CatBoost and LightGBM, are the best performing ML models for this problem; and improving logistic models with SHAP interpretation tools can achieve similar performance levels to the best performing ML methods.

II. Data Description and Preprocessing

The Panel Study of Income Dynamics (PSID) [9] is the longest-running national level longitudinal panel survey of American families. Due to its panel structure and long history, PSID data has become an important data source for life course research [10], [11]. Since 1999, PSID has started to collect individual vehicle information biennially for up to three vehicles in each family that together covers 95% of the total number of vehicles reported by the families. The public data of vehicle information includes body type, model year, owned or leased, acquisition year, manufacturer, and make. We limit our study to the survey waves from 2003 to 2017, because they include consistent questions about vehicle information (vehicle attributes are summarised in Table I).

The outcome variable of interest in this study is the transaction decision for individual vehicles in the family's existing fleet, whether it is disposed without replacement (termed as "disposed" hereafter), disposed with replacement (termed as "replaced" hereafter), or kept in the family in the next wave. We first create life trajectories of individual vehicles by identifying them from wave to wave. Then the transaction outcome, whether to dispose, replace, or keep an existing vehicle can be determined by comparing the household fleet status between the two adjacent waves.

The 34 input features we include are time varying attributes processed from the longitudinal PSID data, including vehicle attributes, household characteristics, life events and change variables, such as marriage, change of income and employment.

TABLE I
VEHICLE-LEVEL DATA SUMMARY

Vehicle-level Summary	**Population**	**By Vehicle Outcome**		
	Mean	**Kept**	**Disposed**	**Replaced**
Vehicle vintage	9.55	8.95	11.63	9.98
Years in family	6.06	6.25	6.12	5.7
Owned (leased = 0)	0.95	0.97	0.95	0.91
Vehicle body type				
Car	0.54	0.52	0.59	0.55
Pickup	0.15	0.16	0.14	0.13
SUV	0.24	0.25	0.2	0.24
Van	0.07	0.07	0.07	0.08
Vehicle transaction outcome				
Dispose	0.1	0	1	0
Keep	0.59	1	0	0
Replace	0.31	0	0	1
Number of Observations	69,697	40,884	7,178	21,635

III. METHODS

A. Machine Learning Models

Four machine learning models are evaluated in this study.

1) Random Forest: This algorithm [12] builds an ensemble of decision trees, or tree predictors, which depend on randomly and independently sampled vectors over the same distribution. The strength, correlation and monitor error are closely followed to track the growing features in response to the branches splitting.

2) Catboost and LightGBM: Standard gradient boosting methods solve over-fitting problems, but inefficiently. In an effort to make gradient tree boosting more flexible and scalable, Chen [13] created the scalable XGBoost algorithm. XGBoost employs a new regularization technique, instead of optimizing the loss function, to minimize over-fitting. This tactic allows XGBoost to be faster and more robust during tuning. Because the majority of input features are categorical variables, we employ the two gradient boosting based methods, CatBoost and LightGBM, that were shown to have better performance for categorical data [14]. Both these methods are extensions of XGBoost. CatBoost focuses on categorical columns using permutation techniques and target-based statistics [15]. The light gradient boosting machine (LightGBM) further improves standard gradient boosting methods. Microsoft developed LightGBM by growing the decision trees leaf-wise, allowing it to support GPU learning speed, with faster training time, better accuracy, and for larger data [16].

3) Neural Network - Multilayer Perceptron: One of the simplest multi-layers neural network architectures, the multilayer perceptron (MLP) [17], is a hierarchical structure of layers containing individual artificial neurons. The power of MLPs comes from their ability to learn patterns in the training data and to relate them to the output. Mathematically, MLPs are considered universal approximators, which means they are capable of learning any mapping function. The MLP architecture consists of an input layer, at one or more hidden layers and an output layer. Each neuron in the hidden layer receives input from the input layer and fires according to the neuron's activation function. During the forward pass, the output of each layer is passed to the next layer and usually the output layer consists of only one neuron. The error is calculated based on the function to be predicted and the output of the network. After the forward pass, the backpropagation algorithm [18] performs a backward pass to adjust the model's weights and biases. This is repeated for many epochs, and it is called training. After training the resulting model can be used for classification and prediction.

B. Multinomial Logit Model

Multinomial logit (MNL) models are the most widely used choice models, and are based on utility maximization theory for predicting multi-class outcomes. The utility of the "keep" outcome, i.e. no transaction, is fixed at 0 without any loss of generality, while the utility function from choosing transaction outcome $j \in dispose, replace$ for vehicle n during wave t in family i is:

$$U_{njit} = \alpha_j + X'_{nt}\beta_j + Z'_{it}\gamma_j + Year_t \cdot \delta_j + \epsilon_{njit} \quad (1)$$

β_j is the alternative-specific coefficient vector associated with the vehicle attributes X'_{nt}, and γ_j is the alternative-specific coefficient vector associated with the family level attributes Z'_{it}. We include year-specific effects $Year_t \cdot \delta_j$ in our model to account for temporal influences affecting every household in the same year: for instance, the economic recession beginning in 2008. To account for serial correlation across time observations within families, we cluster the standard errors of the estimates at the family level.

The Baseline MNL model uses all the input features without interactions. Then SHAP importance and interaction scores (explained in the next method section) are used to improve the model specification in the Improved MNL model by selecting variables and interaction terms.

C. SHAP Values for Feature and Interaction Selection

Recent advances in ML interpretability include algorithms and methods that are able to consistently rank feature importance and to reason about individual predictions and has been applied to transportation research [19]. One of the

most promising methods is SHAP (SHapley Additive exPlanations) [20], an algorithm based on coalitional game theory and Shapley values, that shows how individual instances are predicted by quantifying how much each feature impacts the prediction. In this game, the players in a coalition are replaced by the feature values of a data instance.

Shapley values indicate features' contribution to the final prediction of individual instances or the entire dataset. The SHAP implementation provides a feature importance method that satisfies two main requirements: additive feature attribution and additive importance. Also, especially for tree-based methods, it computes fast, consistent, and locally accurate Shapley values in low-order polynomial time by leveraging the internal hierarchical structure of the tree models. Shapley values require a summation of terms over all possible feature subsets. However, for tree-based methods this can be calculated based on each leaf in a tree. An example of a SHAP feature importance plot is shown in Figure 2 (a).

When Shapley values are calculated, usually feature attributions are only allocated among the individual input features, one at the time. However, similar to the idea of interaction terms used for regression models, additional insights can be gained by separating interaction effects between features. One way to do this is to consider pairwise interactions and compute a matrix of attribution values representing the impact of all pairs of features on a given model prediction. The extension of interaction effects can be obtained through the more modern Shapley interaction index [5]. An example of SHAP feature interaction importance is shown in Figure 2 (b) and (c).

D. Performance Evaluation Metrics

Ten metrics are used to comprehensively compare various aspects of the performance of both ML and MNL models for predicting multi-class vehicle transaction outcomes. The outcome class-specific metrics such as *Accuracy*, *Recall*, *F1*, and *Specificity* are first computed from the confusion matrices. Then the multi-class overall performance metrics are derived including:

1) *Overall Accuracy for correct classification*, which indicates the fraction of instances that are correctly classified.
2) *Average Accuracy*, which is based on the sum of the one-vs-all matrices, and represents a binary classification task where one class is considered the positive class and the combination of all the other classes make up the negative class.
3) *Macro-averaged metrics*, which includes Macro-precision and Macro-F1, also known as sensitivity or the true positive rate, is calculated by taking the means of per-class precision, recall and F1, respectively.
4) *Micro-averaged metrics*, which is from the sum of the one-vs-all matrices for each class, and the sum of these matrices will always be a symmetric matrix, so micro-precision, -recall and -F1 will be the same.

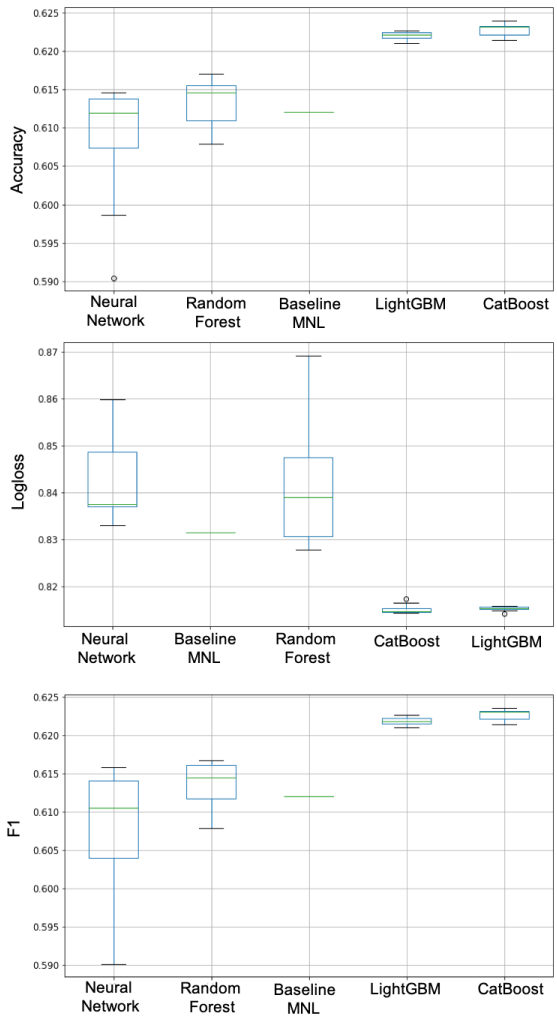
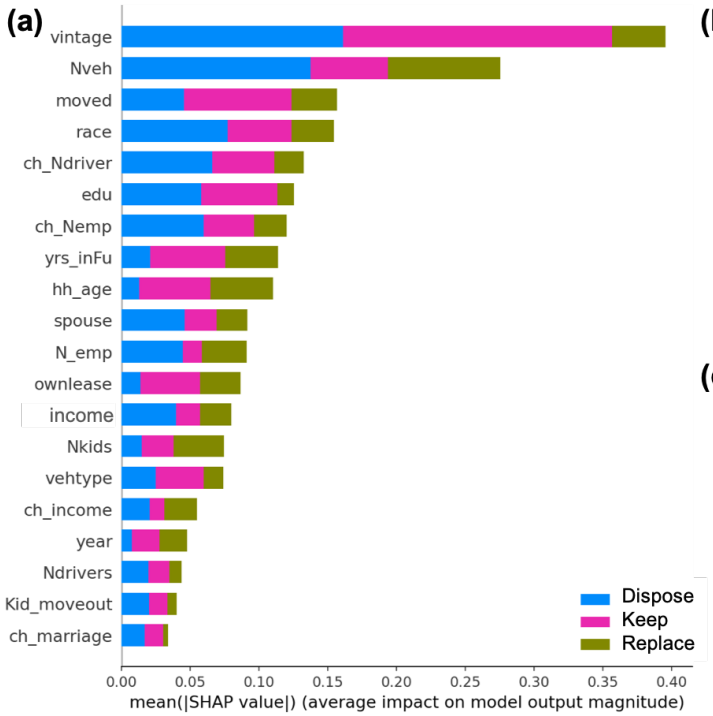


Fig. 1. Performance metrics of four machine learning and baseline MNL models.

Three additional overall performance metrics are computed including:

1) *Cohen's Kappa*, which can be interpreted as a comparison of the overall accuracy to the expected random chance accuracy with higher value indicating a better classifier compared relative to a random chance classifier.
2) *cross-entropy*, which measures the difference between two probability distributions from the idea of entropy in information theory to quantify the number of bits required to transmit an average event from one distribution compared to another with lower cross-entropy as better model performances.
3) *Multi-class Log Loss*, which penalizes the model for uncertainty in correct predictions, and heavily penalizes the model for making an incorrect prediction with lower multi-class log loss as better model performances.



(b)

Interaction Terms Predict "Replace"	SHAP Interaction Score
Nveh:Ndrivers	13.9
vintage:ownlease	10.2
Nveh:vintage	7.7
Nveh:inc_5bins	7.0
Nveh:Nkids	6.9
ch_income:income	6.1
Nveh:ch_income	6.1

(c)

Interaction Terms Predict "Dispose"	SHAP Interaction Score
Nveh:spouse	55.8
ch_Ndriver:Ndrivers	36.1
Nveh:income	34.5
ch_income:income	31.9
Nveh:race	28.7
Nveh:vintage	26.4
Nveh:ch_Ndriver	23.1

Fig. 2. (a) Feature Importance; and interaction term scores for transaction outcome classes "replace" (b), and "dispose" (c).

IV. Results and Discussions

A. Performance Comparison among Machine Learning and Baseline MNL Models

We first compare the performance among the Neural Network, Random Forest, LightGBM, CatBoost, and the Baseline MNL model. The Baseline MNL model refers to the MNL model with linear terms of the input features. Figure 1 summarises the overall performance evaluation on three well-known performance measures: overall accuracy, multi-class log loss and F1. Each point on the plot represents the average performance of a particular model over 5-fold cross-validation. For each ML method, models were generated based on several sets of features and therefore summarised with a boxplot. All the experiments were run using the MLJAR framework [21]. The two gradient boosting based methods, CatBoost and LightGBM, are the best performing ML models and perform better than the Baseline MNL model. More detailed performance comparisons between the Base MNL and best ML model are presented with ten evaluation metrics (the "Base" and "bML" columns) in Table II. Although ML performance is better than the Baseline MNL model both in-sample and out-of-sample, the differences are more pronounced with in-sample and diminish once both models are evaluated on testing data (vehicle transaction data from a random selection of 1000 households).

B. Improving MNL via SHAP Based Variable and Interaction Term Selection

To improve the Baseline MNL model, we use SHAP importance ranking to select the top 20 features and rank their interactions using SHAP interaction scores (Figure 2). Most notably, 5 out of 7 top interactions involve $Nveh$ (number of vehicles owned), indicating that the effects of other input features on the transaction decisions differ among households with different fleet sizes. Accordingly, we revise the Baseline MNL model by (1) segmenting the households by their fleet sizes into one-vehicle families and extra-vehicle-families, and (2) interacting input features with these household segments. Additionally, we add the common interaction term inc_5bins : ch_income (household income levels and change of income between time steps) present in Figure 2(b)(c). This interaction suggests that the effect of income change depends on household income levels.

The performance of the resulting improved MNL model is evaluated with the 10 metrics in Table II column "iMNL". Similar to the Baseline MNL, the improved MNL model shows poorer performance than the ML model on in-sample data. However, when applied to the testing data, 5 out of the 10 metrics have now indicated same or better performance of the improved MNL model than the ML model, and the performance differences are smaller between the improved MNL and ML compared to between the Baseline MNL and ML models. Furthermore, Table II suggests overall MNL models perform more consistently between in-sample and out-of-sample data than ML models.

TABLE II
PERFORMANCE METRICS FOR BASELINE MNL (BASE), IMPROVED MNL (IMNL), AND BEST PERFORMING MACHINE LEARNING (BML). BEST PERFORMING METRICS ARE INDICATED WITH BOLD FACES.

Metrics	**In-sample**			**Testing Sample**		
	Base	**iMNL**	**bML**	**Base**	**iMNL**	**bML**
Overall Accuracy	0.61	0.62	**0.72**	0.61	**0.62**	**0.62**
Average Accuracy	0.74	0.74	**0.81**	0.74	**0.75**	**0.75**
Macro-precision	0.53	0.53	**0.75**	0.53	**0.55**	0.53
Sensitivity	0.42	0.42	**0.58**	0.42	0.43	**0.44**
Macro-F1	0.41	0.42	**0.62**	0.42	0.43	**0.45**
Micro metrics	0.61	0.62	**0.72**	0.61	**0.62**	**0.62**
Cohen's Kappa	0.16	0.17	**0.42**	0.17	0.19	**0.21**
Specificity	0.71	0.72	**0.79**	0.72	0.72	**0.73**
Cross Entropy	1.59	1.57	**1.44**	1.58	1.58	**1.47**
1/(Log Loss)	1.20	1.22	**1.52**	1.20	**1.22**	1.20

V. CONCLUSION

Logistic regression has long been the gold standard for choice modeling in the transportation field. Despite the rising popularity of machine learning in transportation behavior modeling and prediction, few applications exist in predicting household vehicle transactions. To address this research gap, this paper presents a first use-case of ML application to predicting household vehicle transactions by leveraging a newly processed national panel data set. Model performances are reported for four ML models and the traditional MNL model. We find the two gradient boosting based methods, CatBoost and LightGBM, are the best performing ML models. Overall, MNL models perform more consistently between in-sample and out-of-sample than ML models. The SHAP values are useful for screening feature importance and ranking features interactions so that the MNL models can be specified with fewer input features and with important interaction terms. After feature interactions learned from the SHAP tool are used to improve the MNL model specification, the resulting MNL model can also match (if not exceed) the performance of the ML models as evaluated by half of the metrics used.

ACKNOWLEDGMENT

This work was supported by the Office of Advanced Scientific Computing Research and Vehicle Technologies Office of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

[1] S. Van Cranenburgh, S. Wang, A. Vij, F. Pereira, and J. Walker, "Choice modelling in the age of machine learning," *arXiv preprint arXiv:2101.11948*, 2021.

[4] L. Jin, A. Lazar, J. Sears, A. Todd-Blick, A. Sim, K. Wu, H.-C. Yang, and C. A. Spurlock, "Clustering life course to understand the heterogeneous effects of life events, gender, and generation on habitual travel modes," *IEEE Access*, vol. 8, pp. 190 964–190 980, 2020.

[2] Y. Zhang and Y. Xie, "Travel mode choice modeling with support vector machines," *Transportation Research Record*, vol. 2076, no. 1, pp. 141–150, 2008.

[3] C. R. Sekhar, E. Madhu *et al.*, "Mode choice analysis using random forrest decision trees," *Transportation Research Procedia*, vol. 17, pp. 644–652, 2016.

[5] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[6] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[7] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Modeling stated preference for Mobility-on-Demand transit: A comparison of machine learning and logit models," Nov. 2018.

[8] J. J. Levy and A. J. O'Malley, "Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning," *BMC medical research methodology*, vol. 20, no. 1, pp. 1–15, 2020.

[9] (2017) Panel study of income dynamics, public use dataset. produced and distributed by the survey research center, institute for social research, university of michigan, ann arbor, mi. Institute for Social Research, University of Michigan, Ann Arbor, MI. [Online]. Available: https://psidonline.isr.umich.edu/

[10] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, A. Sim, and A. Todd, "Evaluating the effects of missing values and mixed data types on social sequence clustering using t-sne visualization," *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 2, pp. 1–22, 2019.

[11] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, and A. Sim, "Data quality challenges with missing values and mixed types in joint sequence analysis," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 2620–2627.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[14] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset," *International Journal of Computer and Information Engineering*, vol. 13, no. 1, pp. 6–10, 2019.

[15] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[16] E. A. Minastireanu and G. Mesnita, "Light GBM machine learning algorithm to online click fraud detection," *J. Inform. Assur. Cybersecur*, 2019.

[17] J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[18] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," Cornell Aeronautical Lab Inc Buffalo NY, Tech. Rep., 1961.

[19] A. Lazar, A. Ballow, L. Jin, C. A. Spurlock, A. Sim, and K. Wu, "Machine learning for prediction of mid to long term habitual transportation mode use," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 4520–4524.

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[21] A. Plonska and P. Plonski, "Mljar automated machine learning for humans," Nov 2018. [Online]. Available: https://mljar.com/